Jingyuan Zhou
Perspectives on Computational Analysis
Short Paper #4: Proposing an Experiment
Nov. 7th 2016

Experiment proposal

Many Chinese nonprofits promoting gender equality have been sending out brochures to people on the street and giving one-time lectures in universities to stimulate awareness of feminism; however, there are many potential issues with both of these two approaches. People on the streets could throw away these flyers without even reading them, and lectures as an one direction information transferring format may not be effective for invoking thoughts on feminism and could merely change people's opinion. Since these organizations have limited funding, I'm interested in studying what's a better way to educate Chinese people about feminism through an experiment research design. In my personal experience, reading online discussions of feminism has really made me think about different arguments for and against feminism in an everyday life setting. It has also made me realize changes that feminism can make to benefit both men and women. Thus, I have the hypothesis that reading online discussion and discussing people's opinions is a better way to make people think about and eventually accept feminism values.

In my experiment, I would first perform the online survey that I proposed in the last paper where people choose a statement that best describes their understanding of feminism. Then, from the pool of people who chose "I'm not sure what feminism is, but I think the most important ability for a woman is to be a good wife and a good mother" and are willing to participate in my experiment, I will randomly recruit 150 people.

I will randomly split them into three groups so that these groups are roughly similar to each other in sexuality distribution and education level of its subjects. Before the experiment, subjects will all finish a scenario survey of 12 questions. Each question will describe a situation and the subject will have three choices. One choice will correspond to feminism values, one will be the opposite of feminism and the third is a mild answer that's between the two extremes. For each question, the subject will get one point if his/her choice corresponds to feminism value, minus one point if it's the opposite and zero point if it's the mild answer.

After we record the pre-treatment score of each individual, people of the first group will get a standard lecture on feminism that's in the same format as those previously given by nonprofits. People in the second group will read several threads of online discussions on feminism together and discuss their opinions on these arguments with an instructor. The third group is the control group and people in this group are free to do anything they want. After a duration of two hours, all subjects will get 15 minutes break to relax and clear their mind. They will then do the same scenario survey again. The difference between the scores of each individual is calculated. I'll first calculate the mean of change in scores of people in the control group. Then, for people of the other two groups, their "difference score" is their new score minus their old score minus the mean of change of the control group. This is meaningful because when people are asked to answer the same questions twice, they could change their answer due to hesitation. By subtracting this mean from other groups' scores, we can take care of this potential bias. Finally, we will calculate mean and standard deviation of the "difference score" across each group and analyzed through hypothesis testing.

To assess my experiment design, I will examine its validity, ability to measure heterogeneity of treatment effects and causal mechanism. Firstly, I believe that my experiment design satisfy the basic requirement of statistical conclusion validity, internal validity, construct validity and external validity.

Statically, I'll do a hypothesis testing with the null hypothesis that difference of two scores of subjects in the lecture group is the same as the difference of the discussion group. If my p value is small enough for me to reject the null hypothesis and the difference of discussion group is larger than the lecture group, I can conclude that discussing online arguments is a better way to educate people about feminism. Since hypothesis testing and p value are both standard statistical tools for this kind of experiment, I think my statistical analysis of this experiment is done correctly. As for internal validity, since subjects are recruited randomly and backgrounds of subjects are also evaluated as similar between the three groups, randomization is ensured for this experiment. Compared to alternative digital delivery of treatment, my design of bringing all subjects to a lab gives us more control over this process. It's highly likely that people will not pay attention to online lectures, but they are forced to listen or participate in the discussion in the lab. Since in practice the lectures are given in person, a lab setting is also more consistent with the practice. Thus, our delivery of treatment is also valid; however, construct validity is a tricky issue in this case. Since we depend entirely on the scenario survey to measure one person's understanding of feminism, the design of survey questions is crucial to the precision of our experiment. We could consult advice from gender study experts for the survey design and validate our survey by testing it with a group of people with known position on feminism. Externally validity is not a big concern for our experiment because our goal is to test whether a discussion of online arguments is a better approach to educate people on feminism. If our result shows that it is better, nonprofits could use the exact same setting for their future activities. Nevertheless, due to the discussion based property of this approach, people involved in the discussion group could make a big difference on our result. Even one strongly opinionated person could considerable affect other people's perception. Since there will be an instructor to coordinate the session, the ability of generalizing our result to other situations should not be too weak. Thus, apart from inevitable issue of our survey design, our experiment is considerably valid from the four aspects.

The experiment only has very limited ability to measure heterogeneity of treatment effect. As I have explained above, we obtain more control and consistency with our approach in practice from bringing people to the lab to do the experiment as oppose to online video tutorials and discussions. Thus, considering the cost of bringing people into the lab, our scale of number of participants is constrained. In our case of having only fifty people in each group, we can barely make statistically significant assessment on the difference of effect on different people.

To understand the causal mechanism, we could either ask our participants directly or record their actions during the session. In the first approach, we can ask people to fill out an evaluation of our session, and ask them what they like and dislike of our session. We can also ask people which part of the session do they find the most helpful and whether they would prefer the opposite session or not. For example, if a person is in the discussion group, the opposite session will be the lecture session. By asking for direct evaluations, we can understand what they find more helpful. Another way of understanding the causal mechanism is to observe the average duration of people paying attention to the instructor. This requires33 us to record the entire session and observe the amount of time each person is looking at the instructor or taking notes. Since people are not likely to learn new ideas if they are not paying attention to the session, if the discussion session has a higher average than the lecture session, we can explain that discussion is more helpful because people are more engaged in this session. Considering the manpower involved for this approach, the first method of directly asking for evaluation is probably a better method.

In conclusion, my experiment is well suited to answer my question because it is valid and could show causal mechanism by adding an evaluation survey, but it does lack of ability to measure heterogeneity of treatment effect. Since it satisfies my goal of showing nonprofits a potentially better approach to invoke awareness of feminism, its lack of ability to measure heterogeneity of treatment effect is not a concern for me.

It is also a feasible experiment design. Due to the fact that nonprofits can directly benefit from my findings, I can partner with a gender equality nonprofit organization to conduct the experiment. Thus, I can get some funding to conduct the experiment and benefit from their experience on similar tasks.